



JP9319824

Biblio

Page 1

**esp@cenet****DOCUMENT RECOGNIZING METHOD**

Patent Number: JP9319824  
Publication date: 1997-12-12  
Inventor(s): SHIMA YOSHIHIRO; MARUKAWA KATSUMI; SHINJO HIROSHI;  
NAKAJIMA KAZUKI  
Applicant(s):: HITACHI LTD  
Requested Patent:  JP9319824  
Application Number: JP19960136186 19960530  
Priority Number(s):  
IPC Classification: G06K9/20 ; G06F17/60 ; G06F19/00 ; G06K9/00 ; G06K9/72  
EC Classification:  
Equivalents:

**Abstract**

**PROBLEM TO BE SOLVED:** To recognize a variety of documents with high precision and easily correct them by extracting ruled lines and character lines from an image of a registration completion notice, discriminating the kinds of notices by using the structure and line kinds of a ruled-line frame and the recognition result of the character lines, and inspecting the consistency of the recognition result by using a data base and making a correction display.

**SOLUTION:** An image reduction part 201 reduces an image binarized by an image input part 200 and a ruled-like extraction part 203 extracts the ruled lines. A character line extraction part 206, on the other hand, extracts character lines, a frame extraction part 207 finds a ruled-line frame from the extracted ruled-lines, and a character line image extraction part 202 extracts a character line image for character line coordinates in the frame found by an in-frame character line extraction part 209. A character segmentation part 210 segments characters, one by one, in cooperation with a character discrimination part 211 and outputs the character coordinates. An items discrimination part 214 determines which items a character string in the frame of the document corresponds to, through word collating by using an item content dictionary 218 and a content collating part 219 matches the recognition result against words or addresses.

Data supplied from the esp@cenet database - I2

(51)Int.Cl. <sup>6</sup>	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 K 9/20	3 4 0		G 0 6 K 9/20	3 4 0 C 3 4 0 J
G 0 6 F 17/60		9061-5H	9/00	K
		9061-5H	9/72	Z
G 0 6 K 9/00			G 0 6 F 15/21	J

審査請求 未請求 請求項の数13 O L (全 22 頁) 最終頁に続く

(21)出願番号 特願平8-136186  
 (22)出願日 平成8年(1996)5月30日

(71)出願人 000005108  
 株式会社日立製作所  
 東京都千代田区神田駿河台四丁目6番地  
 (72)発明者 嶋 好博  
 東京都国分寺市東恋ヶ窪1丁目280番地  
 株式会社日立製作所中央研究所内  
 (72)発明者 丸川 勝美  
 東京都国分寺市東恋ヶ窪1丁目280番地  
 株式会社日立製作所中央研究所内  
 (72)発明者 新庄 広  
 東京都国分寺市東恋ヶ窪1丁目280番地  
 株式会社日立製作所中央研究所内  
 (74)代理人 弁理士 小川 勝男

最終頁に続く

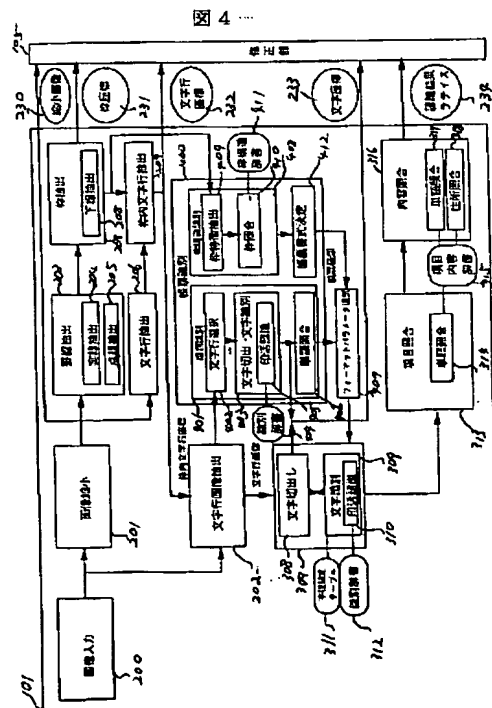
(54)【発明の名称】 帳票認識方法

(57)【要約】

【課題】帳票の種類が多様な読み取り対象に対して、高精度かつ修正が容易な帳票認識手法を提案することである。また、帳票に記載されている罫線のうち、特に、点線を抽出する手法を提案することである。

【解決手段】帳票画像200から罫線と文字行を抽出し、罫線枠の構造と線種並びに認識結果を利用して読み取り項目を決定するとともに、登記簿データベースから該当する登録済の登記情報を読みだし、当該認識結果との照合、検定を行う。

【効果】登記済通知書の名称を読み取り、枠内の文字行を認識するための書式情報を選択的に使用することができるため、読み取り対象が多様な帳票であっても高精度に登記情報を入力できるという効果がある。また、本発明によれば、認識結果の一部を利用して登記情報データベースをアクセスし、当該データベースに具備している登記情報と認識結果の整合性を検定することができるため、登記情報の画面での修正作業が容易であり、効率的な修正作業が可能となる。



**【特許請求の範囲】**

【請求項1】登記済通知書を画像処理し、前記画像から罫線と文字行を抽出し、前記罫線によって構成される罫線枠の構造、線種並びに前記文字行の認識結果の一部を用いることによって前記登記済通知書の種類を識別し、登記情報データベースにアクセスして、前記登記済通知書の認識結果から該当するデータを読み取り、当該認識結果と登記済データベースとの整合性を検定して、該検定結果を基に前記認識結果を修正表示することを特徴とする帳票認識方法。

【請求項2】特許請求の範囲第1項において、上記画像から文字行を抽出し、上記登記済み通知書の最上部にある文字行を上記登記済み通知書の表題として選択し、上記登記済み通知書の種類を単語照合により識別することを特徴とする帳票認識方法。

【請求項3】特許請求の範囲第1項又は第2項において、前記画像から文字行を抽出し、当該文字行から該当する項目を項目内容辞書による単語照合により決定し、予め項目内容辞書に記憶してある項目枠と内容枠との相対的な配置関係により、当該項目枠に対応する内容枠を決定することを特徴とする帳票認識方法。

【請求項4】特許請求の範囲第1項乃至第3項のうちいずれかにおいて、上記罫線のうち点線の抽出は、前記画像を互いに直交するX軸方向とY軸方向に分けたとき、点線とみなせる外接矩形のX軸方向の長さ、Y軸方向の長さ、前記外接矩形の繰り返しピッチ及び前記外接矩形のを直線で結んだと仮定したときの直線の傾きから点線の抽出処理を行うことを特徴とする帳票認識方法。

【請求項5】特許請求の範囲第4項において、前記直線の傾きは、前記画像処理された登記済み通知書から抽出された実線の傾きから得ることを特徴とする帳票認識方法。

【請求項6】特許請求の範囲第4項又は第5項において、上記抽出した点線を構成する外接矩形の繰り返しピッチの規則性を利用して、(1)最も端にある外接矩形を始点として注目し、(2)注目する外接矩形と隣の外接矩形及び一つ隣の外接矩形がそれぞれ存在し、且つ、それぞれ所定の距離内にあるならば、当該外接矩形の隣の外接矩形を終点の候補として注目していくステップを繰り返し、(3)終点の候補として注目する外接矩形の隣の外接矩形が存在し、且つ、当該外接矩形が所定の距離内にあり、一つ隣の外接矩形が存在しないならば、当該外接矩形を終点とし、(4)終点の候補として注目する外接矩形の隣の外接矩形が存在し、且つ、当該外接矩形が所定の距離内にあり、一つ隣の外接矩形が所定の距離内に存在しないならば、当該外接矩形を終点とすることにより点線の始点及び終点を特定することを特徴とする帳票認識方法。

【請求項7】特許請求の範囲第1項乃至第6項のうちいずれかにおいて、前記画像上における罫線と下線とを

区別して抽出し、当該下線が含まれている枠を抽出し、枠内に記載された文字行の読み取り結果を、当該枠内の下線の検出結果によって、有効又は無効とする編集を行うことを特徴とする帳票認識方法。

【請求項8】特許請求の範囲第1項乃至第7項のうち何れかにおいて、前記画像上に点線により区切られた枠に一定の関係がある場合において、前記画像から点線と実線を抽出し、抽出した前記点線の種類によって、前記点線により区切られた二つの枠に関係を設け、当該二つの枠内の文字認識結果を編集出力することを特徴とする帳票認識方法。

【請求項9】登記済通知書の表面画像を入力し文字を読み取る登記情報の認識方法において、登記済通知書の画像から文字行を抽出し、抽出した複数の文字行の位置関係から標題の文字行を選択し、当該標題の文字行を読み取ることにより、登記済通知書の種類を識別することを特徴とする帳票認識方法

【請求項10】登記済通知書の表面画像を入力し文字を読み取る登記情報の認識方法において、予め、点線により区切られて枠の関係を保持しておき、画像から実線と点線を抽出し、抽出した点線により区切られた二つの枠に関係を付け、当該二つの枠内の文字認識結果を編集出力することを特徴とする帳票認識方法

【請求項11】登記済通知書の表面画像を入力し文字を読み取る登記情報の認識方法において、枠罫線と下線とを区別して抽出し、当該下線が含まれている枠を検出し、枠内に記載された文字行の読み取り結果を、当該枠内の下線の検出有無によって、無効又は有効とする編集を行うことを特徴とする帳票認識方法

【請求項12】登記済通知書の表面画像を入力し文字を読み取る登記情報の認識方法において、文字行を抽出し、文字行から該当する項目を単語照合により選択し、予め保持する項目枠と内容枠との相対的な配置関係により、当該項目枠に対応する内容枠を決定することを特徴とする帳票認識方法

【請求項13】登記済通知書の表面画像を入力し文字を読み取る登記情報の認識方法において、登記済データを格納したデータベースを予め保持し、登記済通知書に記載された文字に対する認識結果の一部を用い、当該データベースをアクセスして該当する登記済のデータを読み取り、当該認識結果と登記済データとの整合性を検定して、検定結果を基に認識結果の修正表示することを特徴とする帳票認識方法

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は不動産に関する登記情報が記載された多様な帳票、特に、登記済通知書から文字データを読み取り自動的に入力する帳票認識方法に関する。

【0002】

【従来の技術】従来の帳票認識方法では、読み取り位置が予め決められている定形帳票を認識対象としており、読み取り位置が帳票ごとに異なる登記済通知書を認識することは困難であった。また、登記情報のうち、建物名や面積などの読み取り項目は、辞書用の単語の収集が困難という理由や数字には拘束性が希薄であること理由から、単語照合など後処理により修正を加えることが困難であり、高精度化がむづかしい読み取り項目とされていた。

【0003】帳票中の罫線は、文字が記載された読み取り領域を示す重要な情報源となる。このため、帳票の書式の識別のため、あるいは、読み取り領域の抽出のために、帳票から罫線を抽出することが一般に行われている。また、図面の入力のため、図面に記載された線図形、例えば、実線や破線、鎖線を抽出することが、昭和57年度電子通信学会総合全国大会、1295、P5-280、嶋田茂、角本繁、浅川岩雄「ヒストグラム分類手法による破線・鎖線の認識」において論じられている。罫線の種類としては、一般に、実線、点線、破線、鎖線があるが、本発明で対象とする登記済通知書には、罫線として実線と点線とが印刷されており、これら実線と点線の線種を識別して抽出する必要がある。

【0004】罫線を抽出する方式を画像のデータ形式から大別すると、アイエーピーアールワークショップ オン マシン ビジョン アプリケーション、1992年、163頁から166頁(MVA'92 IAPR WORKSHOP ON MACHINE VISION APPLICATIONS, pp.163-166, Dec. 1992)に論じられている(1)ドットイメージの画素データに対して投影を行いその分布のピークから罫線を抽出する手法(画素投影手法)、昭和57年度電子通信学会総合全国大会、1295、P5-280に論じられている

(2)画像内の塊状図形を長さ方向を持った座標データとしてベクトル化した線セグメントを用いる手法(線セグメント手法)、(3)画像内の塊状図形をランレングス符号化した各黒ランの特徴点を用いる手法(黒ラン手法)、(4)画像内の塊状図形の外接矩形の寸法や位置を用いる手法(外接矩形手法)、が提案されている。

【0005】(1)の画素投影手法は、黒画素を縦及び横方向に投影しその分布のピークから文字行や罫線を抽出する。(2)の線セグメント手法は、図面の認識において、切断された短い線分からなる破線や鎖線を抽出するために用いられている。(3)の黒ラン手法は、ランレングス符号化した黒ランから罫線を抽出する手法である。長い黒ランを基に実線を抽出できることがよく知られている。また、黒ランの中点を2次元平面上にプロットしたドットイメージを用いて、文書中から点線、破線、鎖線やデザイン的な側面の強いフィールドセパレータを抽出する手法が電子情報通信学会論文誌D-2, Vol. J78-D-2, No. 12, pp. 1935-1939(1995年12月)、後藤英昭、阿曾弘具、

「文書画像中のけい線・フィールドセパレータの抽出のための一手法」に論じられている。(4)の外接矩形手法は連結成分の外接矩形を求め、外接矩形の配置を手掛かりに罫線や文字等の文書構成要素を抽出する手法である。外接矩形のうち、内部を占める黒画素の割合の大きいものを点線候補矩形として、その規則的な配置を手掛かりとして点線を抽出する手法が、1994年電子情報通信学会秋季大会D-316, 324頁、園田浩一郎、米田政明、長谷博行、酒井充、「文書画像中の点線抽出アルゴリズムの検討」において論じられている。

【0006】

【発明が解決しようとする課題】本発明の認識対象である登記済通知書はその種類が6種(一般土地、区分所有家屋、一般家屋それぞれ権利分、表示分)あり、また、同じ種類の登記済通知書であっても、記載項目が異なっている。このため、読み取り位置を予め設定しておくことができない非定形帳票であり、従来の帳票認識方法では読み取ることは困難である。本発明の第1の目的は、帳票の種類が多様な読み取り対象に対して、高精度かつ修正が容易な帳票認識手法を提案することである。

【0007】また、罫線を抽出する方法に関して、上記

(1)の画素投影手法は、黒画素を縦及び横方向に投影しその分布のピークから文字行や罫線を抽出する。帳票が傾いている場合、接近した文字行や罫線では分布のピークが検出できないという問題がある。

【0008】上記(2)の線セグメント手法では線セグメントをそのベクトル方向に延長し結合しており、点線のように塊状図形として円形状の点や正方形の点から構成される罫線では、線セグメントの方向を求めることが困難であり、従って、この線セグメント手法では点線を抽出することはできない。

【0009】上記(3)の黒ラン手法では、特徴点を再度ドットイメージに展開し、このドットイメージに対して射影を行い罫線を抽出する手法と同様の画素レベルの処理を含んでおり、処理時間が画素数に応じて長くなるという問題がある。

【0010】上記(4)の外接矩形手法では傾きのある文書から点線を抽出することが考慮されておらず、傾いた点線を抽出することはできない。また、縦横の長さの比が大きい外接矩形を罫線や囲み罫線とみなして分離する手法が提案されているが、点線では外接矩形が複数個に分かれてしまうため、単純には点線を抽出することは困難である。本発明の第2の目的は、罫線のうち、特に、点線を抽出する手法を提案することである。

【0011】

【課題を解決するための手段】上記第1の目的を達成するため、帳票画像から罫線と文字行を抽出し、罫線枠の構造と線種並びに認識結果を利用して読み取り項目を決定するとともに、登記簿データベースから該当する登録済の登記情報を読みだし、当該認識結果との照合、検定

を行う。

【0012】上記第2の目的を達成するため、帳票画像から、まず実線を抽出し、当該実線から帳票の傾きを求める。また、連結成分の外接矩形を抽出する。ついで、外接矩形のうち、横幅と縦幅が所定範囲内の矩形を点線の候補として選択し、この矩形群の内、直線上に配置された外接矩形を先に求めた傾きを考慮して選択する。そして、これら外接矩形の配置の規則性を利用して、点線の始点、終点を検出する。

#### 【0013】

【発明の実施の形態】図1は本発明の一実施例である登記情報認識システムの構成図である。登記情報の認識を行う認識部101と認識結果の修正を行う修正部105がネットワーク104により接続されており、入力センタ111において認識と修正を並行して行うことができる。処理の過程は先ず、スキャナ102により登記済通知書100の画像を入力する。次いで、認識用計算機103では文字および罫線の認識を行い、修正用計算機106でその認識結果の修正確認を行う。また、辞書やコード表と照合チェックし、コードデータを出力する。認識結果は、通信制御用計算機107を介して、遠隔地にある計算センタ110にあるホスト計算機108に接続された登記情報データベース109に格納される。修正部105では、認識結果の一部を利用し、登記情報データベース109をアクセスし登録済の登記情報を読み出す。当該読み出した登録情報と認識結果の一部を照合し、矛盾がないかどうかの検定を行う。

【0014】図2は登記情報認識の処理過程を示すブロック図である。認識部101では、帳票画像を読み取り、修正部105に縮小画像230、枠座標231、文字行座標232、文字座標233、認識結果ラテイス234を送出する。修正部105では、これらの入力データをもとに、操作者が認識結果を修正する。画像入力部200では、帳票表面の画像を白黒2値化して採取する。採取した画像は、画像縮小部201と文字行画像抽出部202に送出される。画像縮小部201では、後続の処理の高速化のため、帳票画像を縮小し、縮小画像230を出力する。縮小処理は、細い罫線が縮小後かすれないよう、画素ごとのOR処理を行う。縮小した画像に対し、罫線抽出部203において罫線を抽出する。ここには、実線抽出部204と点線抽出部205があり、実線抽出部204では黒画素の連続するつながりを基に、縦実線と横実線を抽出する。また、点線抽出部205では連結成分の外接矩形の配置、サイズの拘束条件を用い、縦点線と横点線を抽出する。

【0015】一方、文字行抽出部206では、文字の集合である文字行を抽出する。ここでは、連結成分の外接矩形の頂点座標を基に、外接矩形を融合し、文字行を生成する。枠抽出部207では、抽出した罫線から、罫線が四方を取り囲む枠を求め、枠の頂点座標231を出力

する。この時、下線抽出部208で、文字行の下方に印刷された下線を抽出する。下線は実線であり、始点と終点が他の罫線に接続していないという特徴を利用して下線を抽出する。ただし、下線が短い場合、文字の一部と混同されるため、長い下線を含む枠を求め、当該長い下線を含む枠に対して、所定の位置関係、例えば、左右に隣接する小さい枠にも下線が記載されているとする。枠内文字行抽出部209では、先に文字行抽出部206で抽出した文字行と枠抽出部208で抽出した枠座標を基に、枠内の文字行座標232を求める。この時、複数の枠にまたがって誤って抽出された文字行は、枠罫線によって切断することができる。枠と枠内の文字行が対応つけられて文字行座標が出力される。枠内の文字行座標は文字行画像抽出部202に入力され当該入力画像と文字行座標を基に、文字行画像が抽出され、出力される。文字切出し部210では、文字識別部211と協調しながら文字を1文字ごとに切り出し、その文字座標233を出力する。文字識別部211は、文字パターンを入力として、文字コードを出力する。当該文字識別部211には印活認識部212と識別辞書213が具備されている。

【0016】本発明は、印活認識に限定するものではなく、帳票内の手書き文字を読み取る場合には、手書き文字認識部を具備すればよい。文字識別部211の出力である認識結果は、項目識別部214に入力される。帳票は項目とその項目に対応する内容で構成されており、それらが一般的には枠内に記載されている。例えば、項目として「登記の目的」があり、内容として「抵当権設定」「仮差し押さえ」等がある。項目の名称は予め、単語辞書として項目内容辞書218に格納しておく。項目識別部214では帳票の枠内に記載された文字列がどの項目に該当するかを単語照合により決定する。当該項目識別部214は、字種限定部215、単語照合部216、項目内容対応付け部217からなる。単語照合部216では、項目内容辞書218を用い、項目として記載される単語群と認識結果との照合を行い、項目名称を決定する。項目内容対応付け部217では、項目が記載された枠と、その項目に対応する内容が記載された枠とを決定する。このため、予め、項目の枠と当該枠との相対位置関係を項目内容辞書218に格納している。内容照合部219は、字種限定部220、単語照合部221、住所照合部222からなり、認識結果と単語あるいは住所との照合を行う。これにより、認識結果の精度が向上するという効果がある。

【0017】なお、内容照合では、項目が先に決定しており、内容に関する記載事項に限定がある。この特徴を利用して、字種限定部220では、認識結果から該当する字種以外を除去している。また、住所が記載されていることが項目により分かるため、該当する内容枠では住所照合222を実行する。

【0018】図3は帳票の種類名称(標題)の読み取りを用いた登記情報認識の処理過程を示すブロック図である。図2のブロック図と同じく、認識部101では、帳票画像を読み取り、修正部105に縮小画像230、枠座標231、文字行座標232、文字座標233、認識結果ラテイス234を送出する。画像入力部200では、帳票表面の画像を白黒2値化して採取する。採取した画像は、画像縮小部201と文字行画像抽出部202に送出される。縮小した画像に対し、罫線抽出部203において罫線を抽出する。ここでは、実線抽出部204と点線抽出部205があり、縦実線と横実線、縦点線と横点線を抽出する。

【0019】一方、文字行抽出部206では、文字の集合である文字行を抽出する。枠抽出部207では、枠の頂点座標231を出力する。この時、下線抽出部208で、文字行の下線を抽出する。枠内文字行抽出部209では、枠内の文字行座標232を求める。枠内の文字行座標は文字行画像抽出部202に入力され、文字行画像が抽出され、出力される。帳票識別部300では帳票の種類名称を認識する。帳票の種類名称として、「権利に関する建物登記済通知書(専有)」「権利に関する建物登記済通知書(一般)」「権利に関する土地登記済通知書」等があり、これら帳票種類の名称をよみとり、他の文字行を認識するための文字サイズ、文字ピッチ等のフォーマットパラメータを選択する。当該帳票識別部300において、まず、文字行選択部301で帳票の種類名称が記載されている文字行を、文字行座標をもとに選択する。選択された文字行は文字切出・文字識別部302に送られ、文字切り出しと文字識別を行う。

【0020】本実施例では印活認識部303と識別辞書304を具備している。次いで、単語照合部305で帳票の種類名称を示す文字列との照合を行い、帳票の種類名称を決定する。当該帳票種類は、フォーマットパラメータ選択部に入力され、予め、帳票の種類ごとに格納されているフォーマットパラメータを選択し、出力する。文字切出し・文字識別部307では、当該フォーマットパラメータを用い、文字行画像の認識を行う。文字切出し・文字識別部307には、文字切出し部308と印活認識部310を有する文字識別309が具備されており、字種限定テーブル311と識別辞書312を用い、文字切り出しと文字識別を行う。文字識別部309の出力である認識結果は、項目照合部313に入力される。項目の名称は予め、単語辞書として項目内容辞書315に格納しておく。項目照合部313では帳票の枠内に記載された文字列がどの項目に該当するかを単語照合部314により決定する。項目内容辞書315を用い、項目として記載される単語群と認識結果との照合を行い、項目名称とその項目に対応する内容が記載された枠とを決定する。このため、予め、項目の枠と当該枠との相対位置関係を項目内容辞書315に格納している。内容照合

部316は、単語照合部317、住所照合部318からなり、認識結果と単語あるいは住所との照合を行う。これにより、認識結果の精度が向上するという効果がある。なお、内容照合では、項目が先に決定しており、内容に関する記載事項に限定がある。この特徴を利用し、認識結果から該当する字種以外を除去してもよい。

【0021】図4は帳票の枠の構造並びに帳票の種類の種類名称(標題)を用い、登記情報を認識する処理過程を示すブロック図である。図2のブロック図と同じく、認識部101では、帳票画像を読み取り、修正部105に縮小画像230、枠座標231、文字行座標232、文字座標233、認識結果ラテイス234を送出する。画像入力部200では、帳票表面の画像を白黒2値化して採取する。採取した画像は、画像縮小部201と文字行画像抽出部202に送出される。縮小した画像に対し、罫線抽出部203において罫線を抽出する。ここでは、実線抽出部204と点線抽出部205があり、縦実線と横実線、縦点線と横点線を抽出する。

【0022】一方、文字行抽出部206では、文字の集合である文字行を抽出する。枠抽出部207では、枠の頂点座標231を出力する。この時、下線抽出部208で、文字行の下線を抽出する。枠内文字行抽出部209では、枠内の文字行座標232を求める。枠内の文字行座標は文字行画像抽出部202に入力され、文字行画像が抽出され、出力される。帳票識別部400では帳票の種類名称(標題)を認識するとともに、枠構造を識別する。標題識別部401は図3と同じく、文字行選択部402、文字切出・文字識別部403、印活認識部405と識別辞書404、単語照合部406を具備している。また、枠構造識別部408では、枠特徴抽出部409において、枠の特徴、例えば、枠の相対的な配置関係を抽出し、枠照合部410において予め格納している枠構造辞書411との照合を行い、その結果により、帳票書式決定部412で帳票の書式を選択、生成し、フォーマットパラメータ選択部407に送出する。

【0023】上述したように、当該帳票種類と枠の書式情報は、フォーマットパラメータ選択部407に入力され、予め、帳票の種類ごとに格納されているフォーマットパラメータを選択し、出力する。図3と同じく、文字切出し・文字識別部307では、当該フォーマットパラメータを用い、文字行画像の認識を行う。文字切出し・文字識別部307には、文字切出し部308と印活認識部310を有する文字識別309が具備されており、字種限定テーブル311と識別辞書312を用い、文字切り出しと文字識別を行う。文字識別部309の出力である認識結果は、項目照合部313に入力される。項目の名称は予め、単語辞書として項目内容辞書315に格納しておく。項目照合部313では帳票の枠内に記載された文字列がどの項目に該当するかを単語照合部314により決定する。項目内容辞書315を用い、項目として

記載される単語群と認識結果との照合を行い、項目名称とその項目に対応する内容が記載された枠とを決定する。このため、予め、項目の枠と当該枠との相対位置関係を項目内容辞書315に格納している。内容照合部316は、単語照合部317、住所照合部318からなり、認識結果と単語あるいは住所との照合を行う。

【0024】図5は図2で示した登記情報認識の処理フローを示す図である。ステップ500で帳票の表面画像を入力し、ステップ501で当該画像を縮小する。次いで、ステップ502で画像から罫線を抽出し、ステップ503で罫線から枠を抽出する。また、ステップ504で当該縮小画像から文字行を抽出し、文字行の座標を出力する。ステップ505では、文字行座標と枠座標を用い枠内文字行を抽出する。ステップ506では文字行内の画像を当該枠内文字行の座標をもとに抽出する。ステップ507では文字行画像から文字の切り出しを行い、切り出したパターンに対してステップ508で文字識別を実行し、文字認識結果を出力する。ステップ509で文字認識結果を用い、項目を識別する。そして、ステップ510で当該項目に対応する内容枠の文字認識結果に対して照合を行い、ステップ511でそ項目と内容に対応つけて出力する。

【0025】図6は認識対象である登記済通知書の画像を説明のために簡略的に示した図である。帳票画像600の例では、縦実線601、602、604と横実線605、606、609が印刷されている。また、縦点線603、横点線607、608が印刷されている。項目として面積610と所有者611がある。項目610に対応する枠には、数字列「12」612と数字列「34」613が記載されており、縦点線603により区切られている。この縦点線603は、面積の小数点の位置を示しており、面積の値が12.34である。所有者の項目611には、複数の氏名「山田太郎」614、「鈴木一郎」616、「田中二郎」617が記載されており、横点線607、608が複数の内容を記載するために印刷されている。従って、点線と実線の線種を自動的に区別し、小数点位置や、複数内容の検出を行う。また、下線615が、削除内容として用いられている。従って、下線を検出し、当該下線が付与されている内容を無効とする。

【0026】図7は図5で示した登記情報認識の処理の途中結果を示す図である。図7(a)はステップ500で入力した帳票画像700を示しており、同図(b)

(c) (d) (e)はステップ502で抽出した罫線を示す。罫線の内、横実線710、711、712、713と縦実線720、721、722がそれぞれ抽出されている。また、横点線730、731と縦点線740が抽出されている。同図(f)はステップ505で抽出した枠内文字行750、751、752、753、754、755、756を示す。

【0027】図9は図5のステップ509の項目識別の処理過程を説明する図である。ステップ900では後述する800の認識結果ラテイスを入力し、ステップ901で字種を限定する。字種の限定方法は認識結果ラテイスから使用しない字種を除去する方法である。そして、ステップ902で単語照合を行い項目名称を決定する。ステップ903では項目と内容枠の対応付けを行う。

【0028】図8は図9のステップ902の単語照合における処理過程を説明する図である。認識結果ラテイス800は、第n位までの文字識別結果が文字桁の順に並んだデータであり、横方向が文字桁を表わし、縦方向が認識結果の順位を示す。文字桁は801が第1桁、802が第2桁、803が第3桁であり、例では項目「所有者」を示す。認識結果の順位は810が第1位、811が第2位、812が第3位、813が第4位を示す。項目辞書830には項目に該当する単語831、832、833、834、835が予め具備されている。これらの単語を順次読み出し、認識結果ラテイスと文字列の比較を文字列比較部820で行い、類似度合を類似度合算部821で求める。そして、項目決定部822で類似度合が最大となる単語の選択を行う。この例では、項目辞書830にある単語「所有者」832が認識結果ラテイス800での類似度合が最大で、項目として単語「所有者」が決定される。

【0029】図10は図9のステップ903で示した項目と内容枠との対応付けのために用いる対応テーブルの説明図である。対応テーブル1020には、項目データ1000と対応する内容枠の位置データ1010が格納されている。内容枠の位置データは項目の枠に対する相対的な位置関係で表現されており、例えば、面積1001の内容枠位置として右隣1011が登録されている。同じく、所有者1002の内容枠位置として右隣1012が登録されている。この対応テーブル1020を用いることにより、抽出した項目の枠位置を基に内容枠の位置を決定することができる。

【0030】図11は図6の600で示した帳票画像を読み取ったデータの形式を説明する図である。項目と内容はツリー状に格納される。表全体を表わすノード1100に下位には面積ノード1101と所有者ノード1102があり、これらは、項目名称に該当する。面積ノード1101の下位には整数部ノード1103と小数部ノード1104がある。整数部ノードと小数部ノードとの区別は、抽出した縦点線の線種により行う。

【0031】整数部ノード1103の下位に内容ノード1105として、数字列12が自動的に入力される。また、小数部ノード1104の下位に内容ノード1106として、数字列34が入力される。所有者ノード1102の下位には、複数の所有者を示すノード、所有者1ノード1107、所有者2ノード1109、所有者3ノード1111がある。これらのノードには、無効と有効の

情報が、1108、1110、1112で示すように付与されている。この無効と有効の情報は検出した下線を基に各所有者ノードに付与する。また、読み取った所有者の氏名1113、1114、1115が各所有者ノード1107、1109、1111の下位に格納される。

【0032】図12は図2の下線抽出部208の処理フローを説明する図である。ここでは、下線を検出し、下線が含まれる枠と当該枠とを対応付けて登録する。まず、ステップ1200で抽出された枠が尽きるまで、ステップ1201以下の処理を繰り返す。ステップ1201で当該枠の枠座標を入力し、ステップ1202で罫線登録テーブルを初期化する。次いで、ステップ1203で罫線が尽きるまで、ステップ1204以下の処理を繰り返す。ステップ1204で罫線の始点と終点の座標を入力する。そして、ステップ1205が注目する罫線が当該枠内にふくまれるかどうかの判定を座標値を比較することにより行う。もし、枠内に注目する罫線が含まれると判定された場合は、ステップ1206において枠と対応付けて罫線を罫線登録テーブルに登録する。このような処理を繰り返すことにより、罫線登録テーブルに枠と枠内の罫線、すなわち、下線が対応付けられて格納されることになる。

【0033】図13は図3の301で示した帳票識別の処理フローの例である。この帳票識別の処理では、帳票の上部に記載されている帳票名称を読み取り、帳票の種類を決定する。まず、ステップ1300で文字行を入力し、ステップ1301で文字行を選択する。ここでは、帳票の最上部にある文字行を帳票名称が記載された文字行とし、その座標値を基に該当する文字行を選択する。次いで、ステップ1302で文字行画像から文字切出と文字識別を行い、認識結果ラテイスを出力する。ステップ1303で、認識結果ラテイスに対して、予め、単語として格納している帳票名称と単語照合を行い、ステップ1304で帳票の種類を、照合した単語から決定する。

【0034】図14は図4の400で示した枠構造識別の処理フローの例である。ステップ1400で対象とする帳票の枠に対して、その罫線座標を入力し、ステップ1401で当該罫線を符合化する。符合化は縦罫線と横罫線、それぞれに対して、配置の順に番号を付与する。さらに、縦罫線と横罫線の接続関係を生成する。ステップ1402では、枠構造の照合を行い、予め設定している枠構造辞書1403の中から対象とする枠構造ともっとも類似する枠構造を求める。当該枠照合は、特開平7-141462、羽田野英一、児玉和行、嶋好博、古賀昌史、栗野清道、杉本建行、「文書システム」に論じられている手法を用いてもよい。ステップ1404で枠構造から帳票種類を決定し、ステップ1405で枠と項目並びに内容の対応付けを行う。

【0035】図15は図2の105で示した修正部にお

いて登記情報の検定処理の説明図である。1500は登記済通知書の一例であり、項目として「所在」1501、「地番」1502、「地目」1503、「地積」1504、「権利者」1505、「義務者」1506が記載されており、これら項目に対する内容として、「何区何町」1507、「何番何」1508、「宅地」1509、「123」1510、「456」1511、「山田太郎」1512、「鈴木二郎」1513がそれぞれ記載されている。ここで、内容のうち「123」1510は地積の整数部、「456」1511は地積の小数部であり、点線により区切られている。この登記済通知書1500に対して、認識部101において、認識結果1520が得られる。認識結果1520は、表1521の下位に所在1522、地番1523、地目1524、地積1525、権利者1526、義務者1527が項目として抽出されており、それぞれの項目に対応した内容として、それぞれの下位に内容が付与されている。所在1522の下位に何区何町1528、地番1523の下位に何番何1529、地目1524の下位に宅地1530、地積1525の下位に数値「123.456」1531、権利者1526の下位に山田太郎1532、義務者1527の下位に鈴木二郎1533が付与されている。

【0036】一方、図1の109で示した登記情報データベースには、所在地データベース1550と登記データ1560が格納されている。所在地データベースは階層構造をしており、区域の下位に地番がある。不動産に関する登記データ1560は地番の下位に具備されている。本実施例では、区域1551の下位に地番「1番1」1552、「何番何」1553があり、地番「何番何」1553の下位に登記データ1560が具備されている。登記データの一例として、1560に示すように、地目1561の下位に「宅地」1564、地積1562の下位に「123.456」1565、権利者1563の下位に「鈴木二郎」1566が具備されている。検定処理1540では、認識結果1520の所在、地番を基に、所在データベース1550、登記データ1560をアクセスし、地目、地積等の登記情報が認識結果と整合性があるかどうかを判定する。

【0037】図16は修正部105の処理フローを示す図である。ステップ1600で縮小画像230を入力し、ステップ1601で枠座標231を入力する。また、ステップ1602で文字行座標232を入力し、ステップ1603で文字座標233を入力する。さらに、ステップ1604で認識結果ラテイス234を入力する。ステップ1605で、枠座標より該当する枠内画像を切り出し、ステップ1606で枠内画像を表示する。ステップ1607で認識結果の項目「所在」と「地番」に対応する内容を選択し、ステップ1608で登記済の登記情報を登記情報データベース109から入力する。そして、ステップ1609で、図15で説明した登記情



報の検定を行い、ステップ1610で認識結果を項目と内容を対応付けて画面に表示する。ステップ1611で整合性に矛盾がある項目、内容についてステップ1612で該当する内容にリジェクトを付与するとともに、画面にも当該項目、内容にリジェクトであることを表示する。画面にリジェクトを表示することは、当該項目を点滅させる、あるいは、色を付ける、または疑問符等の記号を表示する等により容易に実現できる。

【0038】図17は修正部105における画面の表示例である。画面1700において、項目として、「種類」1701、「所在」1702、「地番」1703、「地目」1704、「地積」1705、「権利者」1706、「義務者」1707が表示されている。認識結果として、各項目に対応する読み取り内容が、1711、1712、1713、1714、1715、1716、1717に表示されている。また、それぞれの内容枠の枠内画像を1722、1723、1724、1725、1726、1727に表示している。ここで、認識結果1715は、先に述べた登記情報の検定1609により、登記データと比較して矛盾があると判断され、疑問符が該当内容に付与されている。操作者は、当該画面1700を目視でチェックし、認識結果をキーボードからの入力により修正する。

【0039】図18は図5のステップ502で示した罫線抽出処理のフローである。処理は先ず、ステップ1800で実線を抽出し、次いで、ステップ1801で点線を抽出する。なお、実線の抽出に関しては、例えば、1992年電子情報通信学会秋季大会、D-218、p. 6-220、古賀昌史、中島和樹、丸川勝美、嶋好博、藤澤浩道、「2値画像中の棒状図形の傾き検出の一手法」並びに、1993年情報処理学会第46回(平成5年前期)全国大会、8C-6、pp. 2-219から2-220(1993年)、古賀昌史、中島和樹、丸川勝美、嶋好博、藤澤浩道、「棒状図形の傾き検出のラン符号による高速化の一手法」に論じられているので、本実施例では説明しない。図19はステップ1801で示した点線抽出の処理フローである。先ず、ステップ1900で水平方向のパラメータを設定し、ステップ1901で横点線を抽出する。次いで、ステップ1902で垂直方向のパラメータを設定し、ステップ1903で縦点線の抽出を行う。

【0040】点線の抽出処理の前提条件は、点のサイズ、点のピッチのパラメータがほぼ指定できることである。予め、登記済通知書の点線のパラメータを設定して、書式情報として点線抽出処理に用いる。点線抽出では点がほぼ一定の間隔で出現し、しかもそれらの点は直線上に配置されていることを利用する。点線は図21に示すように4つのパラメータ、すなわち、点の横幅(wid<sub>th</sub>)、縦幅(height)、点線の傾き角、点の繰り返し周期(pitch)で表現される。点の横幅

と縦幅により点のサイズを決定できる。また、点は傾きを有する直線の上に配置されており、それらの点の繰り返し周期は、ほぼ一定である。点線の傾きは、帳票の傾きから求めることができる。なお、点線抽出では、帳票の実線から傾きを予め求めておくものとする。

【0041】図20はステップ1901、1903で示した点線抽出の全体フローである。点は複数個の黒画素が連結した塊状図形であり、この塊状図形の外接矩形を処理単位として点線の抽出を行う。そのため、先ず、帳票の2値画像から連結成分を抽出し、それらに外接する矩形を求める。外接矩形は、左上の頂点座標と右下の頂点座標で表現し、これら頂点座標を処理の単位とする。次いで、(1)抽出した多数の外接矩形の中から、外接矩形の寸法、周囲条件により点線を構成する点の候補を選択する。ここでは、注目する外接矩形の寸法が所定の範囲内にあるかどうか、また、周囲条件として注目する外接矩形の近傍における他の外接矩形の有無をチェックする。この周囲条件を利用することにより、かすれ文字の一部が点線の点と混同することを防いでいる。さらに、(2)外接矩形を傾き方向に投影する。ここでは、前述したように、傾き角は予め、実線で求めておくものとする。外接矩形の投影値を手掛かりに、直線上に配置している外接矩形を選択する。そして、直線上に配置された当該外接矩形を直線上の並びの順にソートする。次いで、(3)点の繰り返し周期はほぼ一定であることを利用して、点の集まりを生成し、これら点の集まりを一つの点線として登録し、それぞれ始点と終点を抽出する。

【0042】点線抽出の詳細処理過程を(1)外接矩形の寸法、周囲条件による選択、(2)外接矩形の傾き方向への投影による選択、(3)外接矩形のピッチ規則性による点線の始点、終点の抽出、の処理順に説明する。

(4)外接矩形の寸法、周囲条件による選択を説明する。図22は寸法により外接矩形を選択するフローである。処理は外接矩形が尽きるまで以下の処理を繰り返す。即ち、該当する外接矩形の横幅と縦幅を矩形の頂点座標から算出し、そして、横幅が所定範囲に入っているか、さらに、縦幅が所定範囲内に入っているかどうか判定する。もし、横幅、縦幅とも所定の範囲内であれば、当該外接矩形を選択する。外接矩形の寸法は、図23に示すように、横幅が所定値TW<sub>max</sub>とTW<sub>min</sub>の範囲内に、縦幅が所定値TH<sub>max</sub>とTH<sub>min</sub>の範囲内にあるとしている。縮小画像の解像度が100dpi換算で、TW<sub>max</sub>=TH<sub>max</sub>=3ドット、TW<sub>min</sub>=TH<sub>min</sub>=0ドットとしている。

【0043】図24は外接矩形の周囲条件を用いて、点線を構成する点の候補を選択する処理フローである。注目する外接矩形が尽きるまで以下の処理を繰り返す。即ち、先ず、注目する外接矩形の中心座標を算出し、参照する外接矩形が尽きるまで、以下の処理を繰り返す。即

ち、参照外接矩形の横幅と縦幅を算出するとともに、参照外接矩形の中心座標を算出する。参照外接矩形の横幅と縦幅がともに所定範囲外であり、かつ、注目外接矩形と参照外接矩形との中心間距離が所定範囲外であれば、当該注目する外接矩形を選択する。参照する外接矩形は、図25に示すように、その横幅が所定値 $NW_{max}$ と $NW_{min}$ の範囲外に、縦幅が所定値 $NH_{max}$ と $NH_{min}$ の範囲外にあるとしている。さらに、注目外接矩形と参照外接矩形との中心間距離が $R$ 以上であるとしている。なお、縮小画像の解像度が同じく $100\text{dpi}$ 換算で、 $NW_{max}=NH_{max}=20$ ドット、 $NW_{min}=NH_{min}=4$ 、 $R=8$ ドットとしている。

【0044】次いで、(2)外接矩形の傾き方向への投影による選択を説明する。図26に示すように、外接矩形の中心座標 $(x_c, y_c)$ を傾き $a$ の方向に投影し、そのときの切片 $b=y_c - a \cdot x_c$ を算出し、番地 $b$ の投影値の値を1だけ加算する。ここで、 $a$ は回転角であり帳票の傾きを示しており、予め、実線の傾きから求めた値を用いる。外接矩形の投影値Projectionの分布は、図27に示すように、直線の切片の位置にピークを有している。この特徴を利用して、投影値が所定値 $Pr_{th}$ より大きい切片位置を点線候補の切片位置とする。なお、画像の解像度が同じく $100\text{dpi}$ 換算で、横の点線抽出用として $Pr_{th}=40$ 個、縦の点線抽出用として $Pr_{th}=3$ 個、としている。そして、次の処理過程に移り、図28に示すように、求めた点線候補の切片位置 $b$ に対して、 $a+d$ から $a-d$ までの範囲の切片を有する外接矩形を当該点線を構成する点であるとし、それらの外接矩形を選択する。ここで、 $d$ は一定値である。なお、縮小画像の解像度が同じく $100\text{dpi}$ 換算で、 $d=3$ ドットとしている。

【0045】さらに、(3)外接矩形のピッチ規則性による点線の始点、終点の抽出について述べる。点線上にある点の周期はほぼ一定である。外接矩形の中心間の距離を求め、一定の間隔であれば、当該外接矩形は点線を構成する点であるとする。外接矩形の中心間の距離は、図29に示すように、太枠で示した注目外接矩形に対して、隣の外接矩形との距離が所定値 $P1_{max}$ 以内であり、かつ、一つ隣の外接矩形との距離が所定値 $P2_{max}$ 以内であれば、注目外接矩形は点線を構成する点であるとする。なお、縮小画像の解像度が同じく $100\text{dpi}$ 換算で、 $P1_{max}=10$ ドット、 $P2_{max}=20$ ドットとしている。

【0046】図30は直線上に配置されている外接矩形から複数の点線を抽出し、それらの点線の始点と終点を検出する処理の説明図である。直線上に配置された外接矩形をソートしておき、左端から順次、点線の始点と終点を決定していく。先ず、左端の外接矩形を第1番目の点線の始点として登録し、注目する外接矩形を順次左にずらしながら、先に述べたピッチの規則性を注目外接矩形の隣の外接矩形、その隣の外接矩形が保持しているか

判定する。点線の終点の決定は、注目する外接矩形の隣の外接矩形がピッチの条件を満たし、その一つ隣の外接矩形がピッチの条件を満たさない場合、当該注目する外接矩形の隣の外接矩形を終点とする。そして、さらに、次の外接矩形を第2番目の点線の始点として登録し、同じく、終点の外接矩形を探索する。このようにして、直線上に配置されている外接矩形から複数の点線を抽出し、それぞれの点線の始点と終点を検出する。

【0047】

【発明の効果】本発明によれば、登記済通知書の名称を読み取り、枠内の文字行を認識するための書式情報を選択的に使用することができるため、読み取り対象が多様な帳票であっても高精度に登記情報を入力できるという効果がある。

【0048】また、本発明によれば、認識結果の一部を利用して登記情報データベースをアクセスし、当該データベースに具備している登記情報と認識結果の整合性を検定することができるため、登記情報の画面での修正作業が容易であり、効率的な修正作業が可能となる。

【0049】また、本発明によれば、罫線のうち点線を抽出し、点線により区切られた左右の枠が整数部と小数部であることを自動的に判定することができ、登記情報のうち、面積、地積等の数値情報を整数部と小数部を結合させて入力できるという効果がある。

【図面の簡単な説明】

【図1】本発明の一実施例である登記情報認識システムの構成図である。

【図2】登記情報認識の処理過程を示すブロック図である。

【図3】帳票の種類名称の読み取りを用いた登記情報認識の処理過程を示すブロック図である。

【図4】帳票の枠の構造並びに帳票の種類の名義を用いた登記情報認識の処理過程を示すブロック図である。

【図5】図2で示した登記情報認識の処理フローを示す図である。

【図6】認識対象である登記済通知書の画像の説明図である。

【図7】図5で示した登記情報認識の処理の途中結果を示す図である。

【図8】図9のステップ902の単語照合における処理過程を説明する図である。

【図9】図9のステップ903で示した項目と内容枠との対応付けのために用いる対応テーブルの説明図である。

【図10】図9のステップ903で示した項目と内容枠との対応付けのために用いる対応テーブルの説明図である。

【図11】図6の600で示した帳票画像を読み取ったデータの形式を説明する図である。

【図12】図2の下線抽出部208の処理フローを説明

する図である。

【図13】図3の301で示した帳票識別の処理フローの例である。

【図14】図4の400で示した枠構造識別の処理フローの例である。

【図15】図2の105で示した修正部において登記情報の検定処理の説明図である。

【図16】修正部105の処理フローを示す図である。

【図17】修正部105における画面の表示例である。

【図18】図5のステップ502で示した罫線抽出処理のフローである。

【図19】ステップ1801で示した点線抽出の処理フローである。

【図20】ステップ1901、1903で示した点線抽出の全体フローである。

【図21】点線のパラメータを説明する図である。

【図22】寸法により外接矩形を選択するフローである。

【図23】注目する外接矩形の寸法条件を説明する図である。

【図24】外接矩形の周囲条件を用いて、点線を構成す

る点の候補を選択する処理フローである。

【図25】参照する外接矩形の条件を説明する図である。

【図26】外接矩形の傾き方向への投影による外接矩形の選択を説明する図である。

【図27】外接矩形の投影値の分布を説明する図である。

【図28】所定範囲の切片を有する外接矩形を選択する処理を説明する図である。

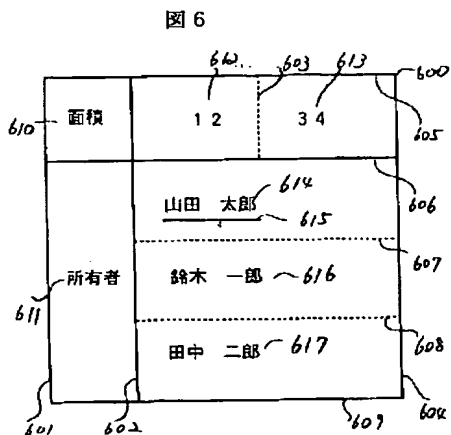
【図29】注目外接矩形が点線を構成する点である条件を説明する図である。

【図30】直線上に配置されている外接矩形から複数の点線を抽出し、それらの点線の始点と終点を検出する処理の説明図である。

【符号の説明】

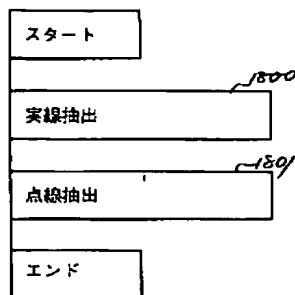
100…登記済通知書、101…認識部、105…修正部、109…登記情報データベース、205…点線抽出部、208…下線抽出部、313…項目照合、316…内容照合、400…帳票識別部、1020…対応テーブル。

【図6】

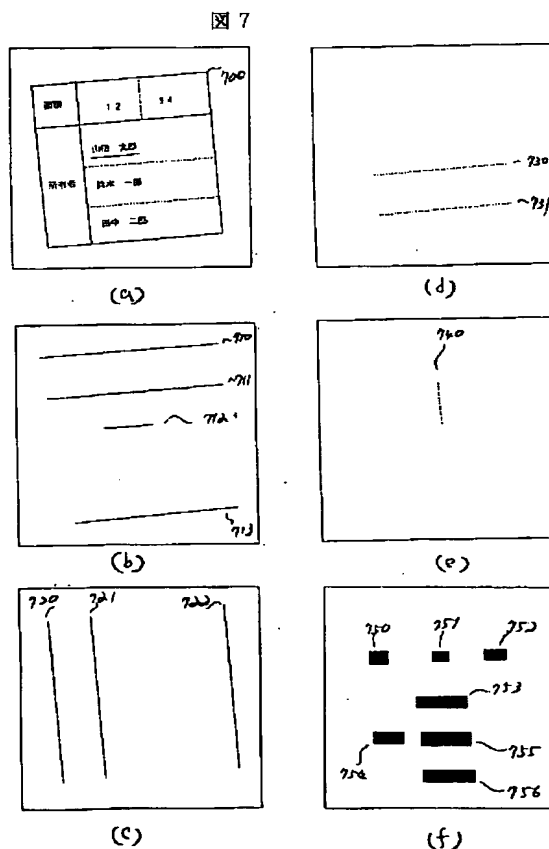


【図18】

図18



【図7】

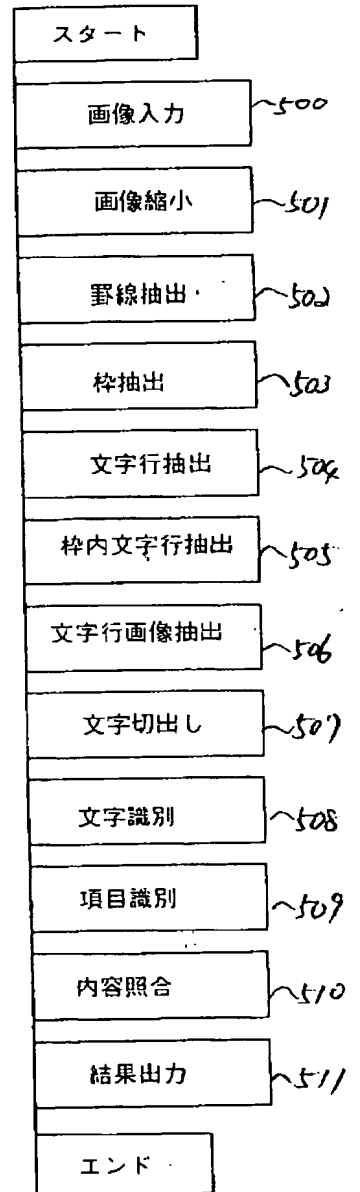
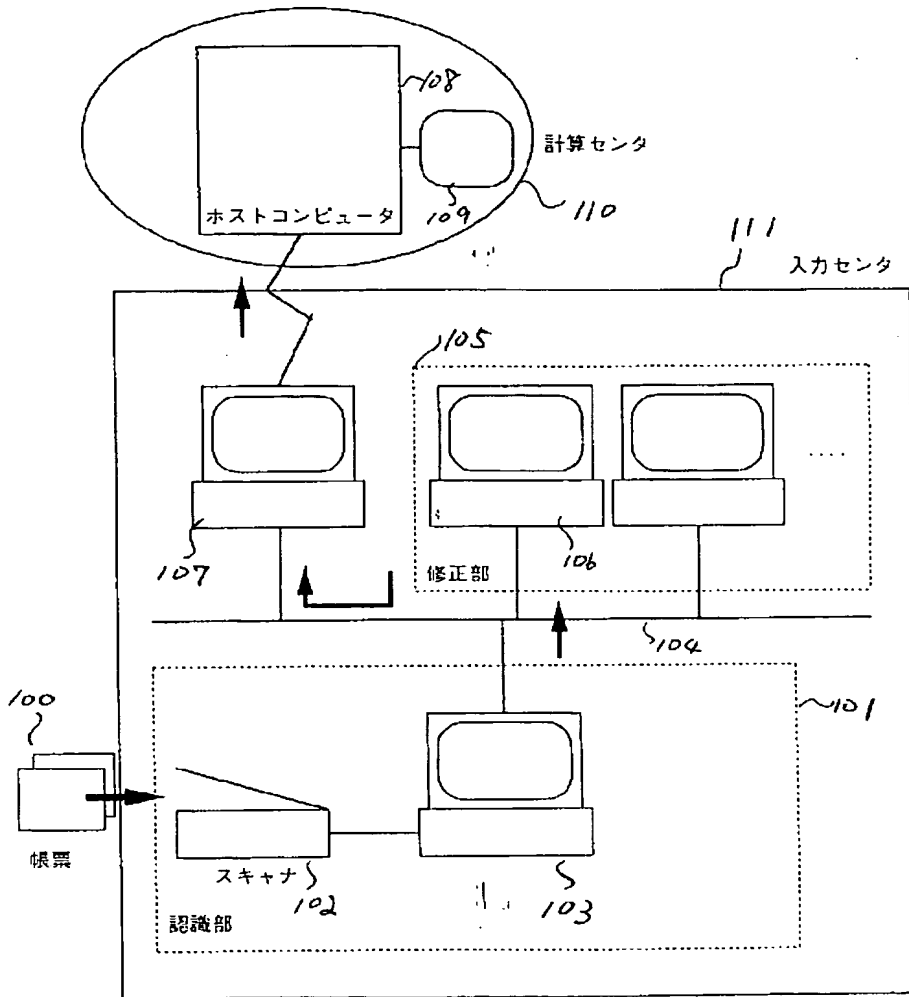


【図1】

【図5】

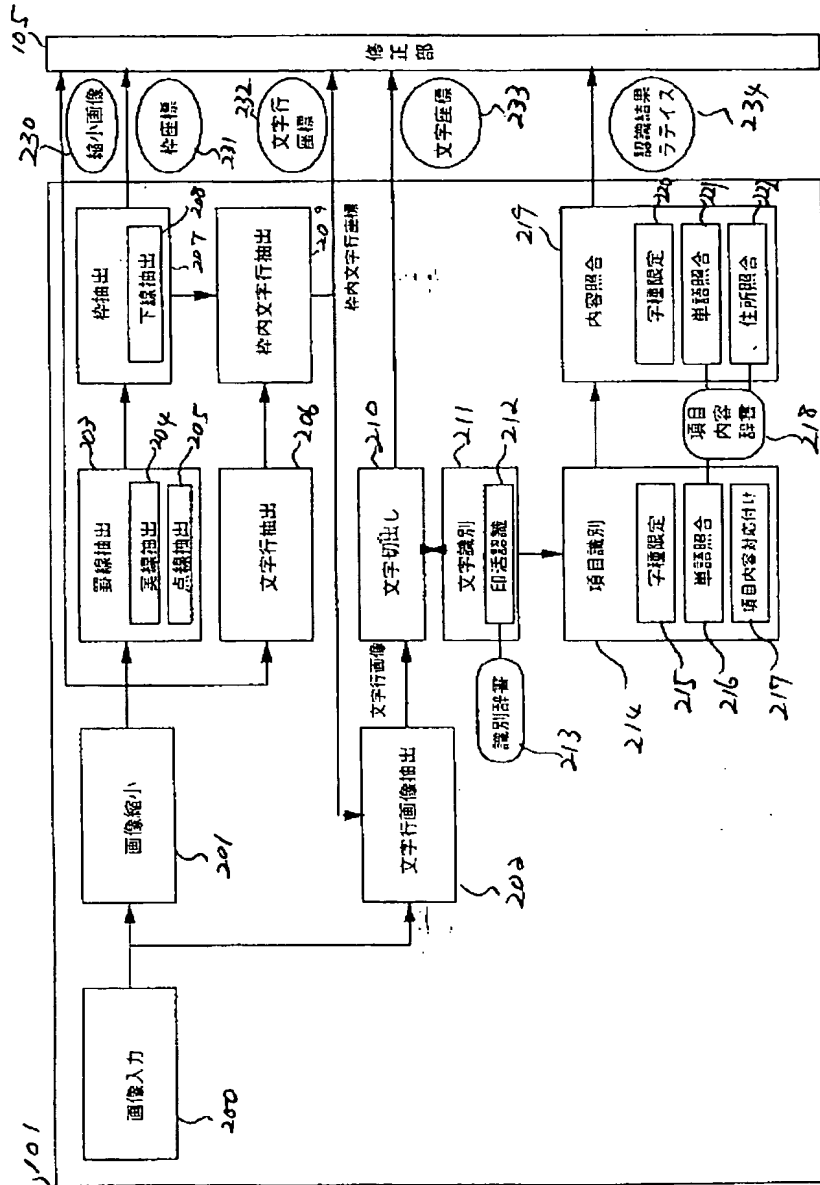
図 1

図 5

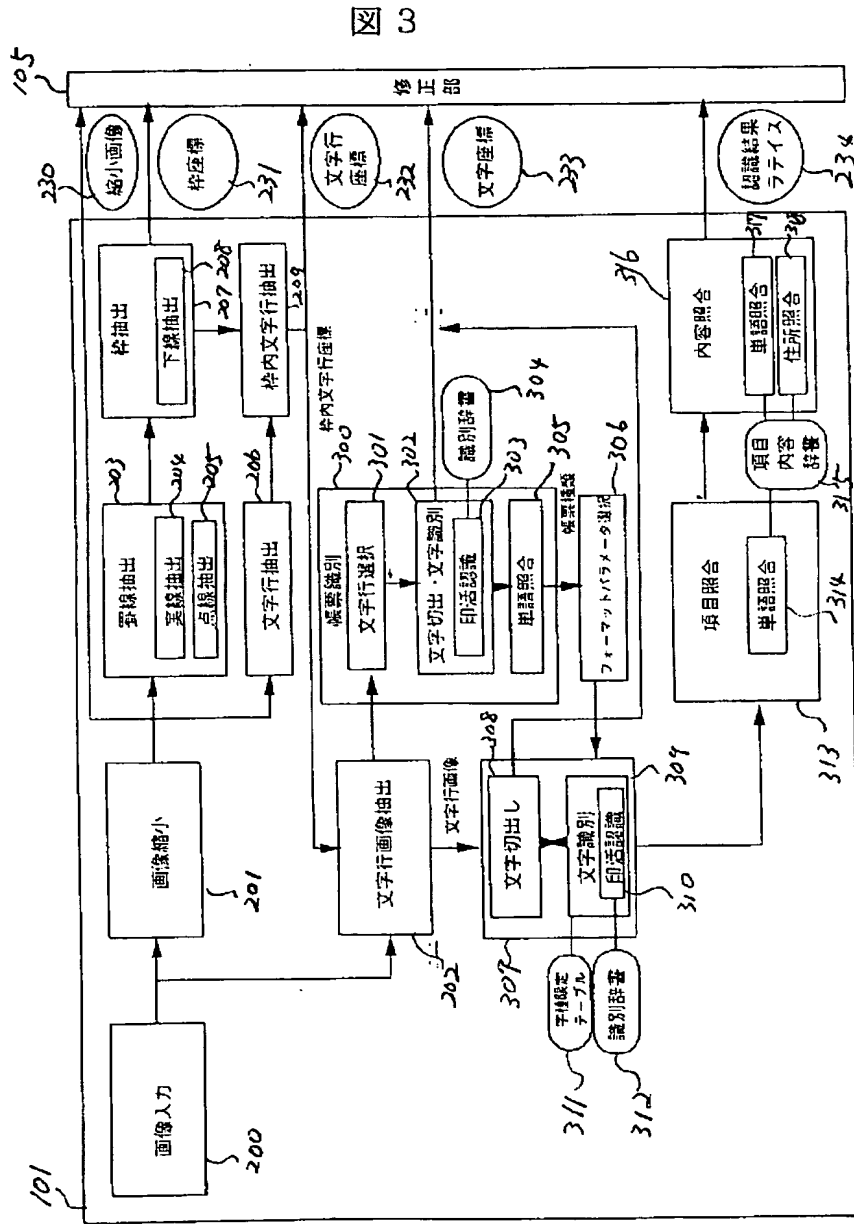


【図2】

図 2



【図3】





【図8】

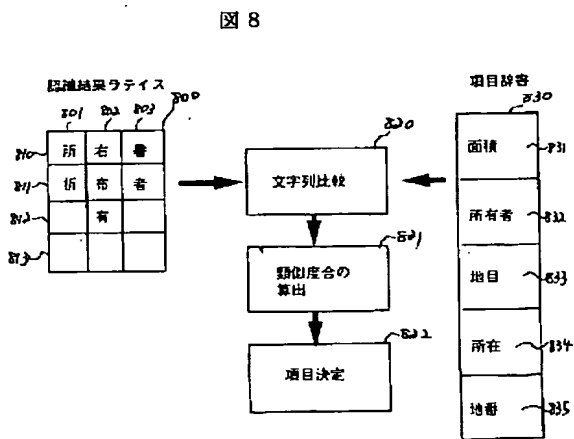


図 8

【図9】

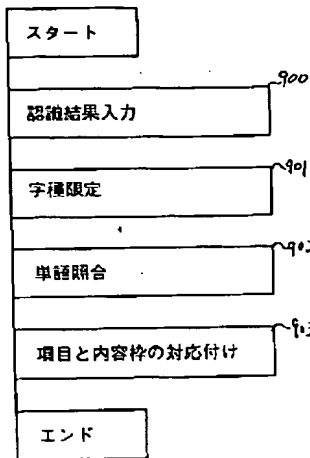


図 9

【図19】

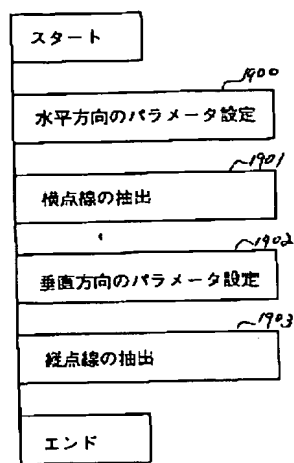


図 19

【図10】

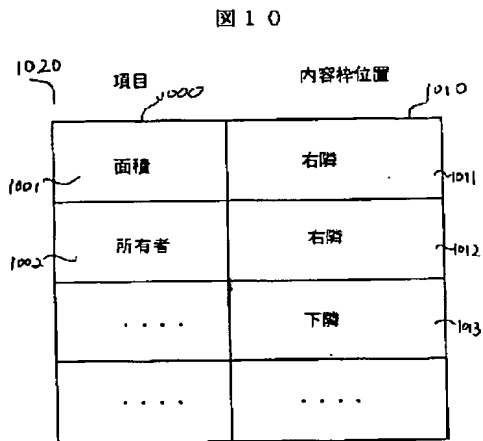


図 10

【図11】

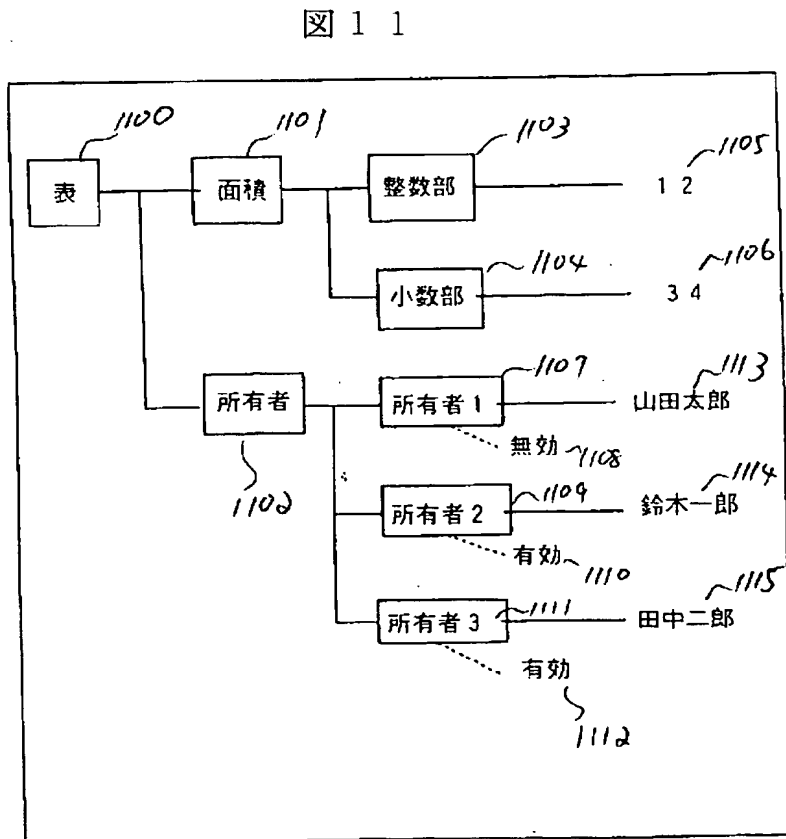


図 1 1

【図23】

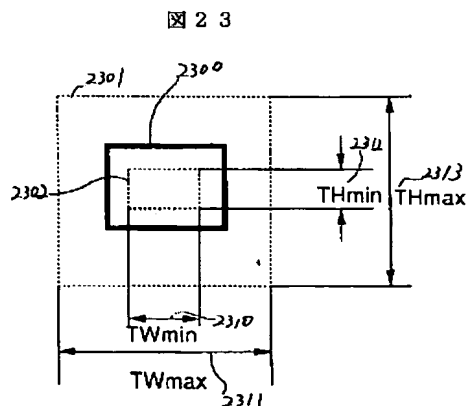
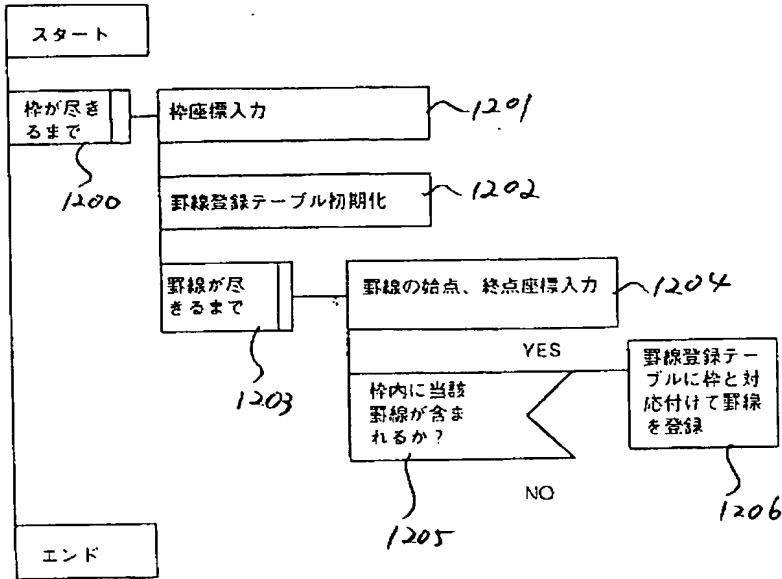


図 2 3



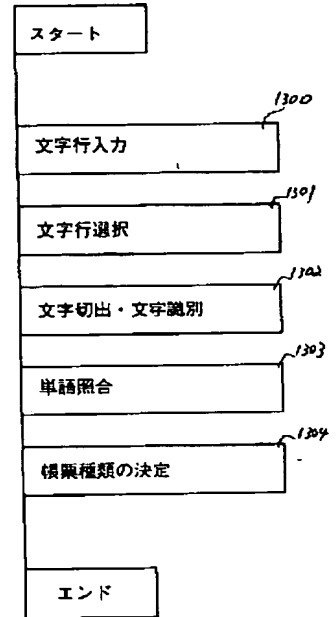
【図12】

図12



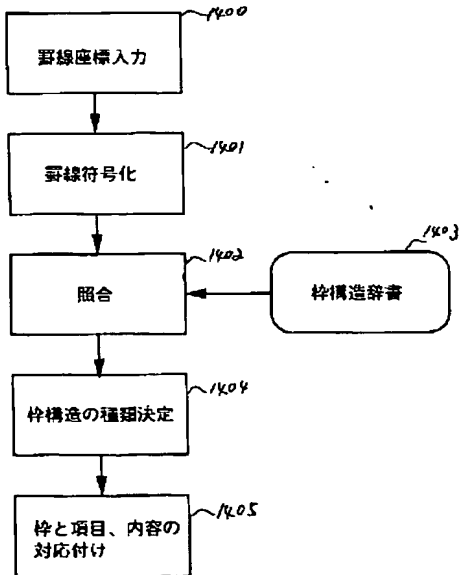
【図13】

図13



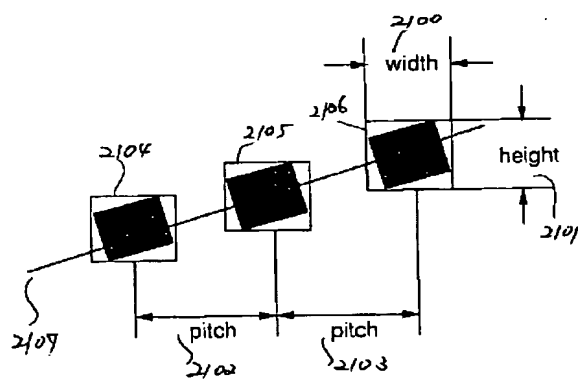
【図14】

図14



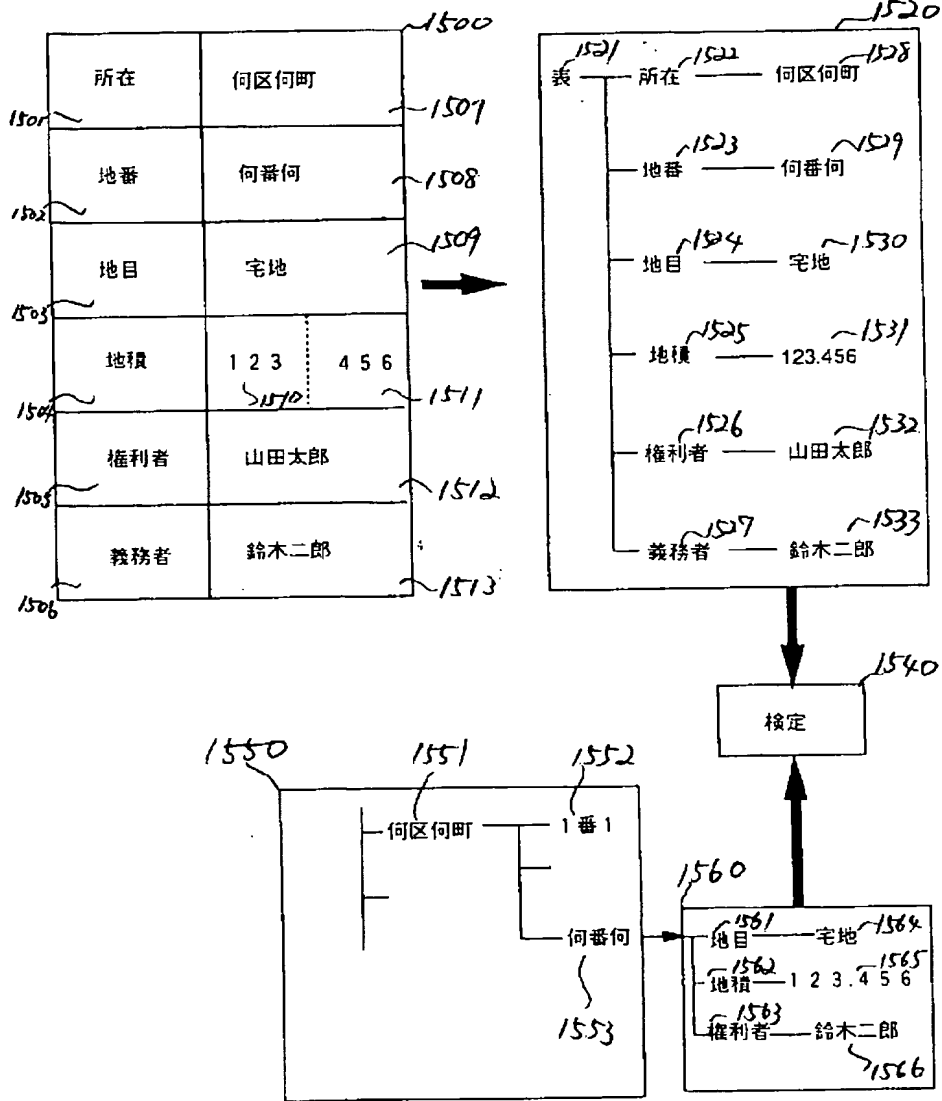
【図21】

図21



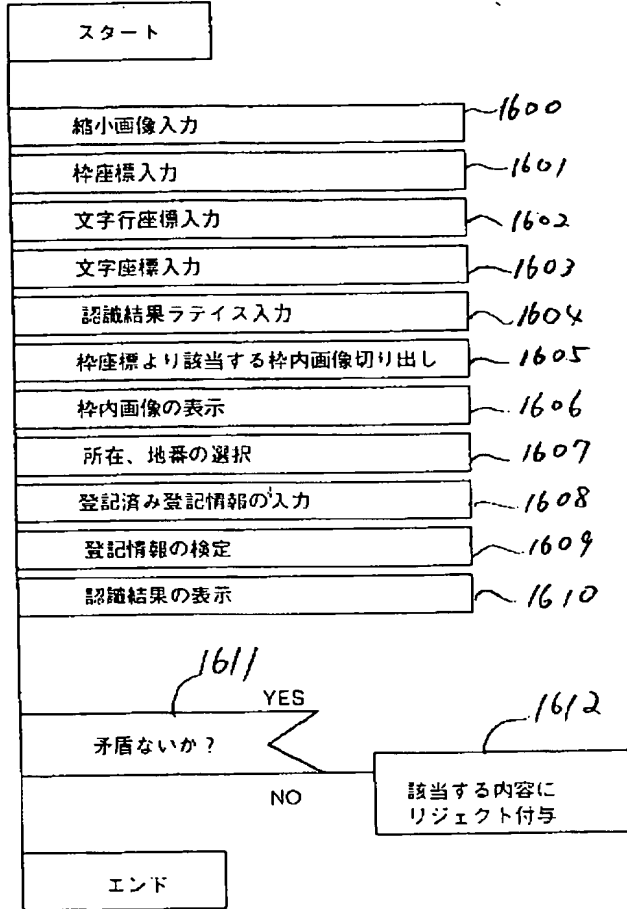
【図15】

図15



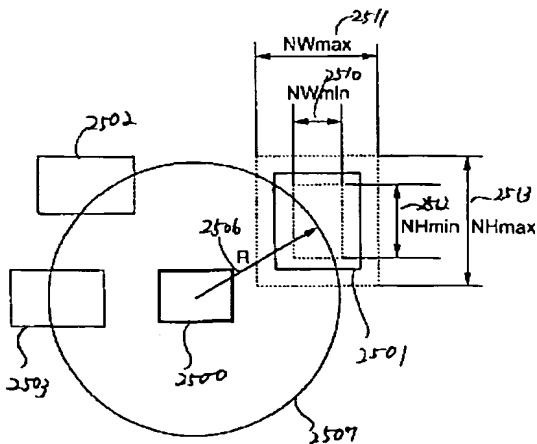
【図16】

図16



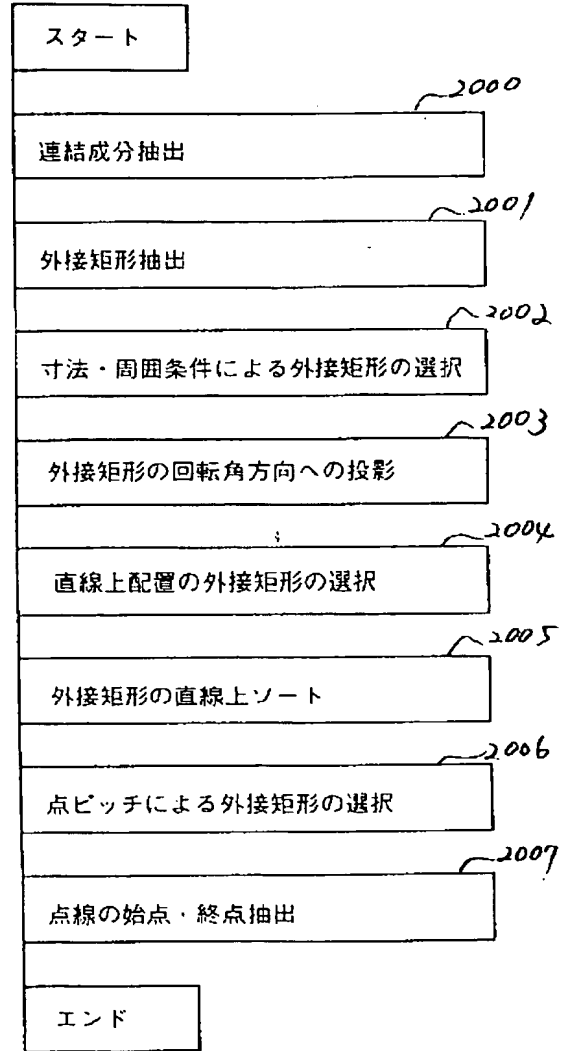
【図25】

図25



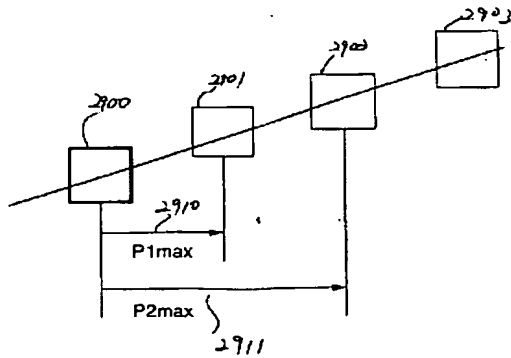
【図20】

図20



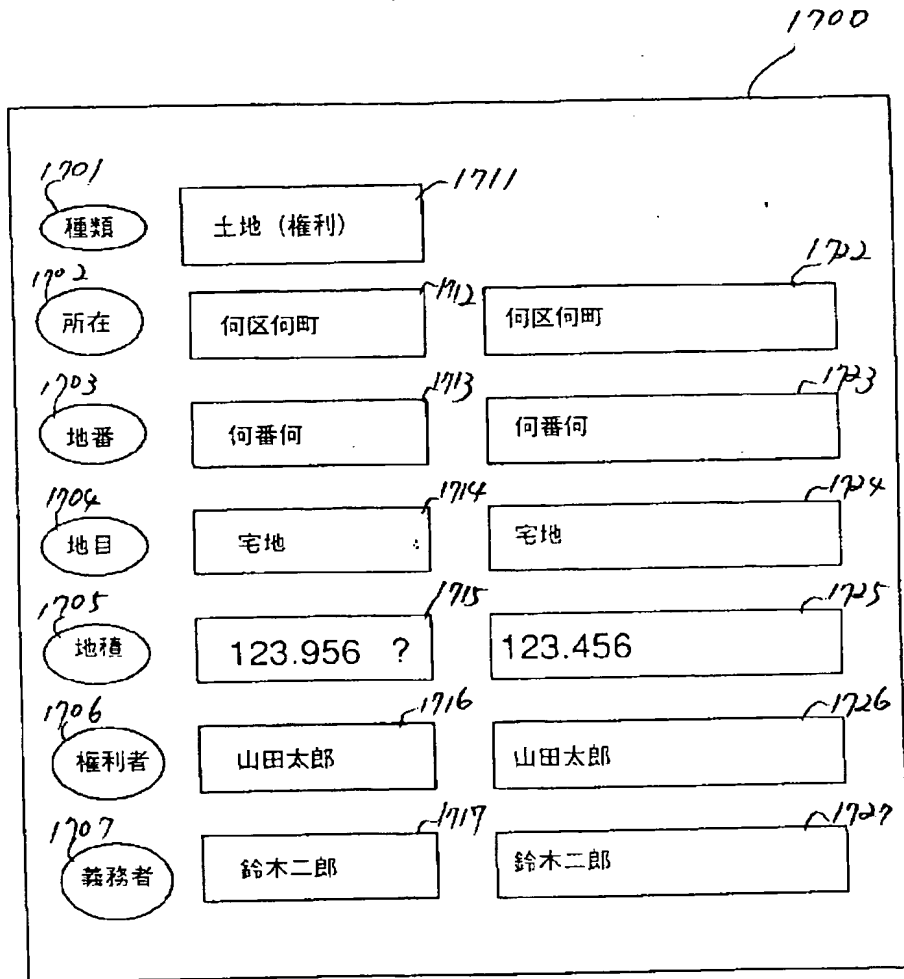
【図29】

図29



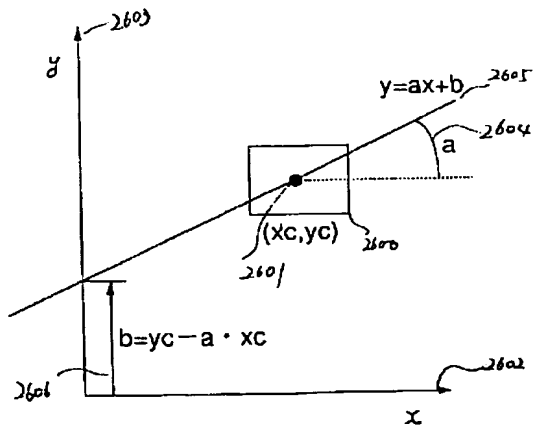
【図17】

図17



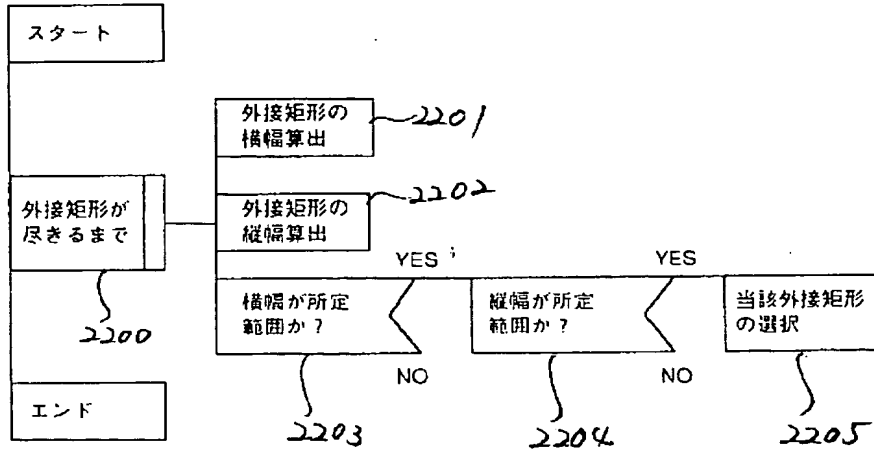
【図26】

図26



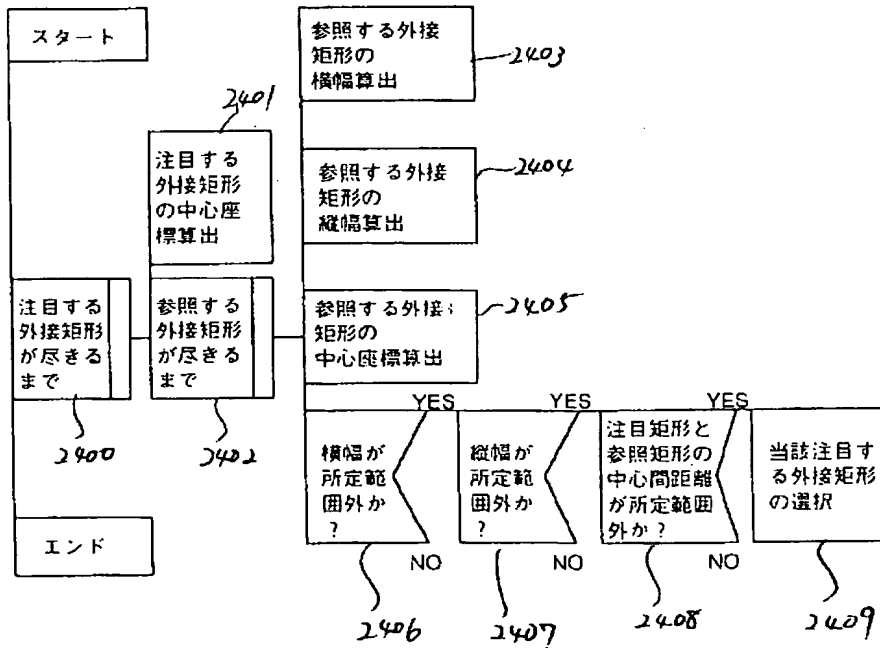
【図22】

図22



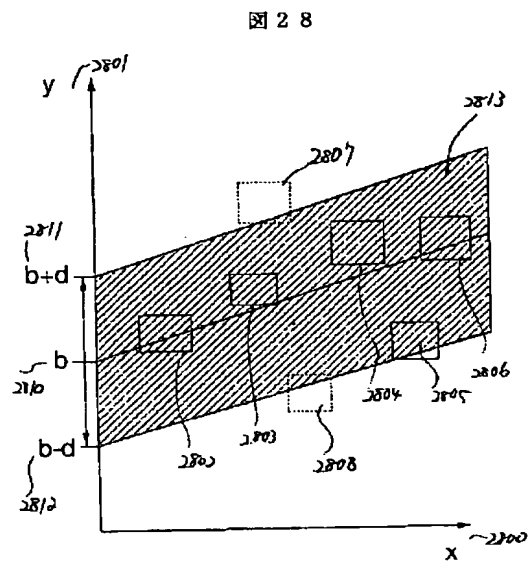
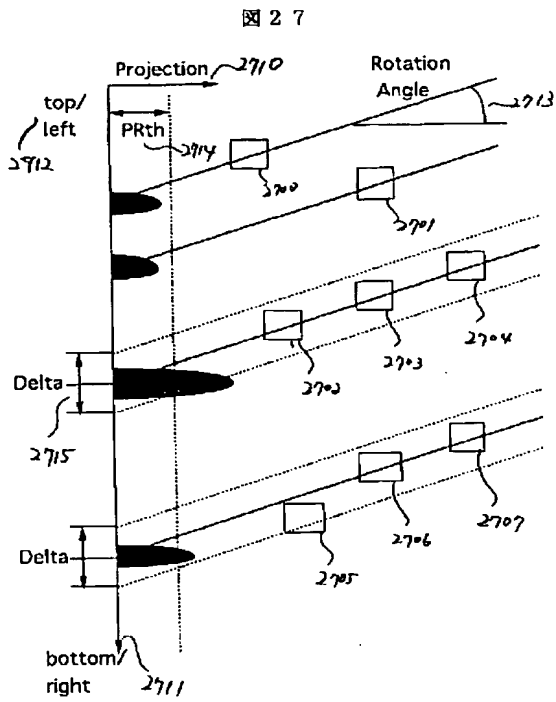
【図24】

図24



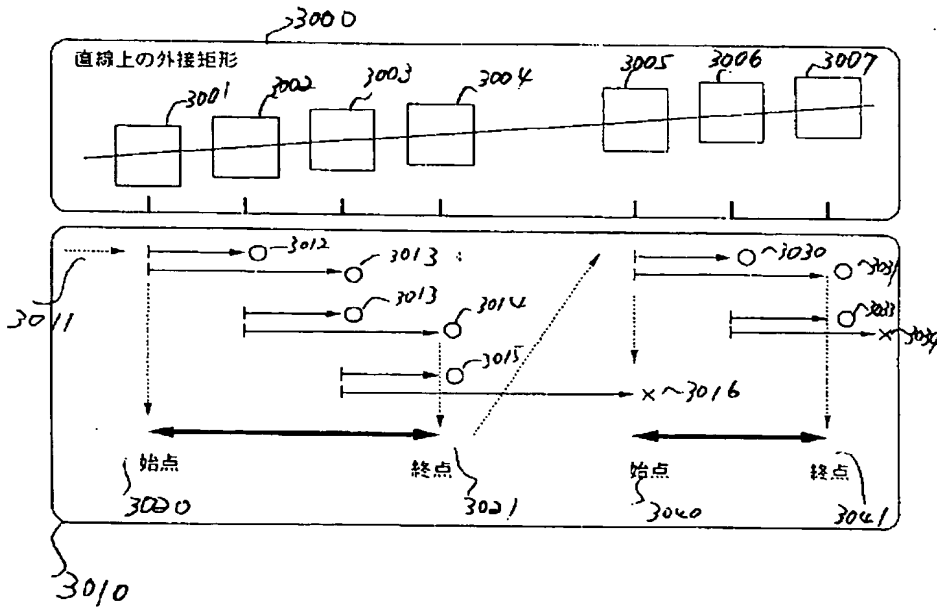
【図27】

【図28】



【図30】

図 30



フロントページの続き

(51) Int. Cl. <sup>6</sup>	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 K 9/72			G 0 6 F 15/22	C G

(72) 発明者 中島 和樹  
東京都国分寺市東恋ヶ窪 1 丁目280番地  
株式会社日立製作所中央研究所内